Genes for Geeks

An introduction to molecular biology for computer scientists.

Humberto Ortiz Zuazaga

humberto@hpcf.upr.edu

http://www.hpcf.upr.edu/~humberto/

Outline

- Why Bioinformatics?
- The "Central Dogma of Molecular Biology" (not!)
- Cells are the stuff of life
- Proteins
- Genes
- Transcription and translation
- Evolution

Bioinformatics

Def: *Bioinformatics* is the application of computer science to biological problems. Computational molecular biology concentrates on the applications of computer science to the study of biological sequences (of amino acids and nucleotides).

The Central Dogma of Molecular Biology (not!)

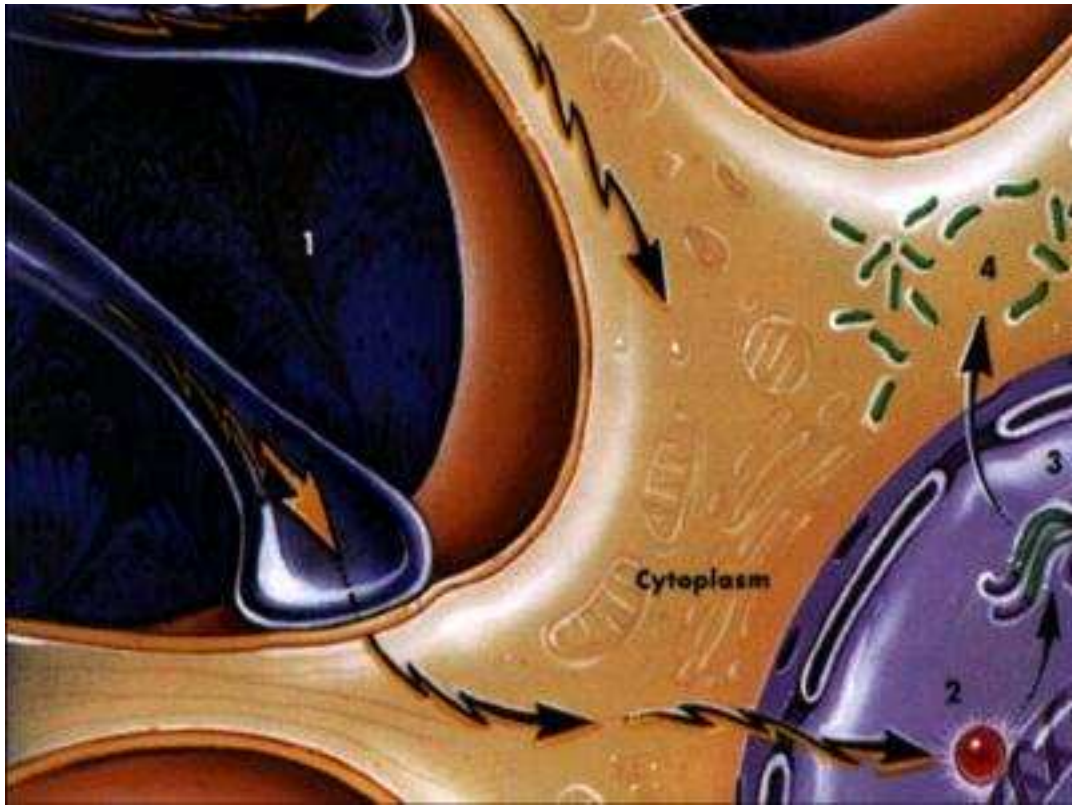
1. Genetic information is stored in our DNA
2. The DNA of a gene is copied to make RNA
3. The RNA of a gene is copied again to make a protein

If we know the complete DNA sequence of an organism, we should be able to predict the complete state of the organism (not!)

The scale of the problem

- The human genome is about 3.2 billion base pairs (3.2 Gb)
- We know about 85% of the human genome
- There are about 35,000 genes
- There are about 100,000 proteins
- Changes in a single base pair are responsible for many illnesses including hereditary breast cancer.

A Model Cell



Modern Molecular Biology in 2 minutes: 1, signals are received at the cell surface, and travel eventually to the nucleus, 2 where transcription factors cause the signal to be converted into a change in expression of a gene. 3 The gene products are converted to proteins in the cytoplasm, 4 where they can now effect further changes in the cell.

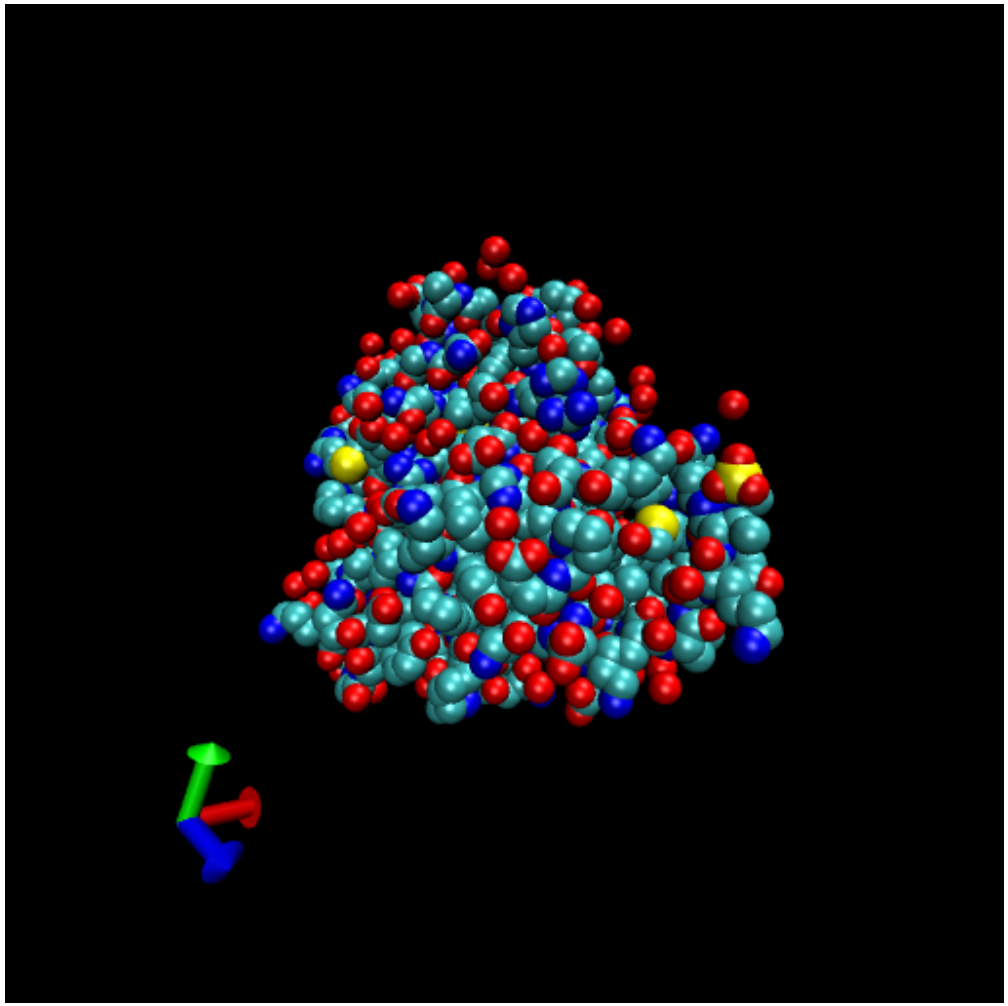
Proteins

- Proteins are most of the components of cells
- Proteins construct the components of cells that are not protiens
- 3D structures composed of one or more polypeptides
- Polypeptides are linear sequences of amino acids that are chained together
- An amino acid is a small organic molecule, there are about 20 different amino acids

Myoglobin

- First protein structure ever determined
- Related to hemoglobin
- Binds and stores oxygen in muscle tissue
- Gives red meat it's color

mcbo



Amino acids

- Building blocks of proteins
- One end has carboxy (COOH) group
- The other end has an amino (NH_2) group
- Amino and carboxy groups separated by a single carbon
- Each amino acid differs in the side chain attached to the carbon

Structural biology

- The function of a protein (what it does) is completeley determined by its structure (3D shape)
- The structure of a protein is completely determined by the sequence of its polypeptide components
- The first biopolymers to be sequenced were proteins, but now it is much simpler, faster, and cheaper to sequence DNA

It should be possible (but in practice it is not) to predict the function of any protein from it's sequence.

Genomics

- Genomic DNA is a linear sequence of 4 nucleotides (A, C, G, T)
- DNA forms the double helix by pairing with its reverse complement (A-T, G-C)
- Genomic DNA contains many genes, each of which is formed from one or more exons (stretches of genomic DNA), separated by introns
- A gene is copied into complementary RNA in a process called transcription (U substitutes T)
- An RNA sequence is transcribed 3 bases at a time (a codon) into an amino acid

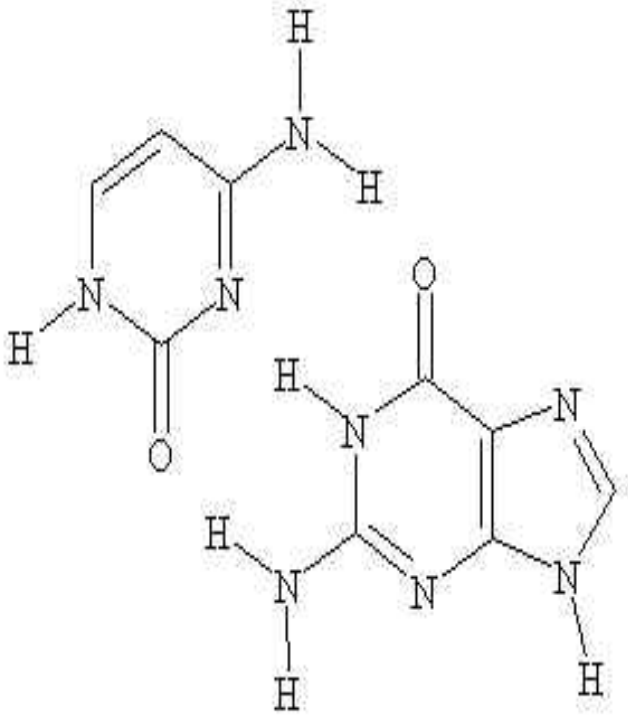
Nucleotides

- DNA and RNA are linear molecules
- Linear structure links sugars with phospate groups at 3' and 5' positions
- Heterocyclic bases different, give each nucleotide it's properties
- One ringed bases are pyrimidines (C, T, U)
- Two ringed bases are purines (A, G)

The famous double helix

- strands of DNA are formed by linking the sugars with phospates
- by convention, read 5' to 3'
- two strands wind around each other, making the famous double helix
- strands in double helix run in opposite directions (antiparallel)
- hydrogen bonds across strands form base pairs and stabilize helix

CG pair



Codon table

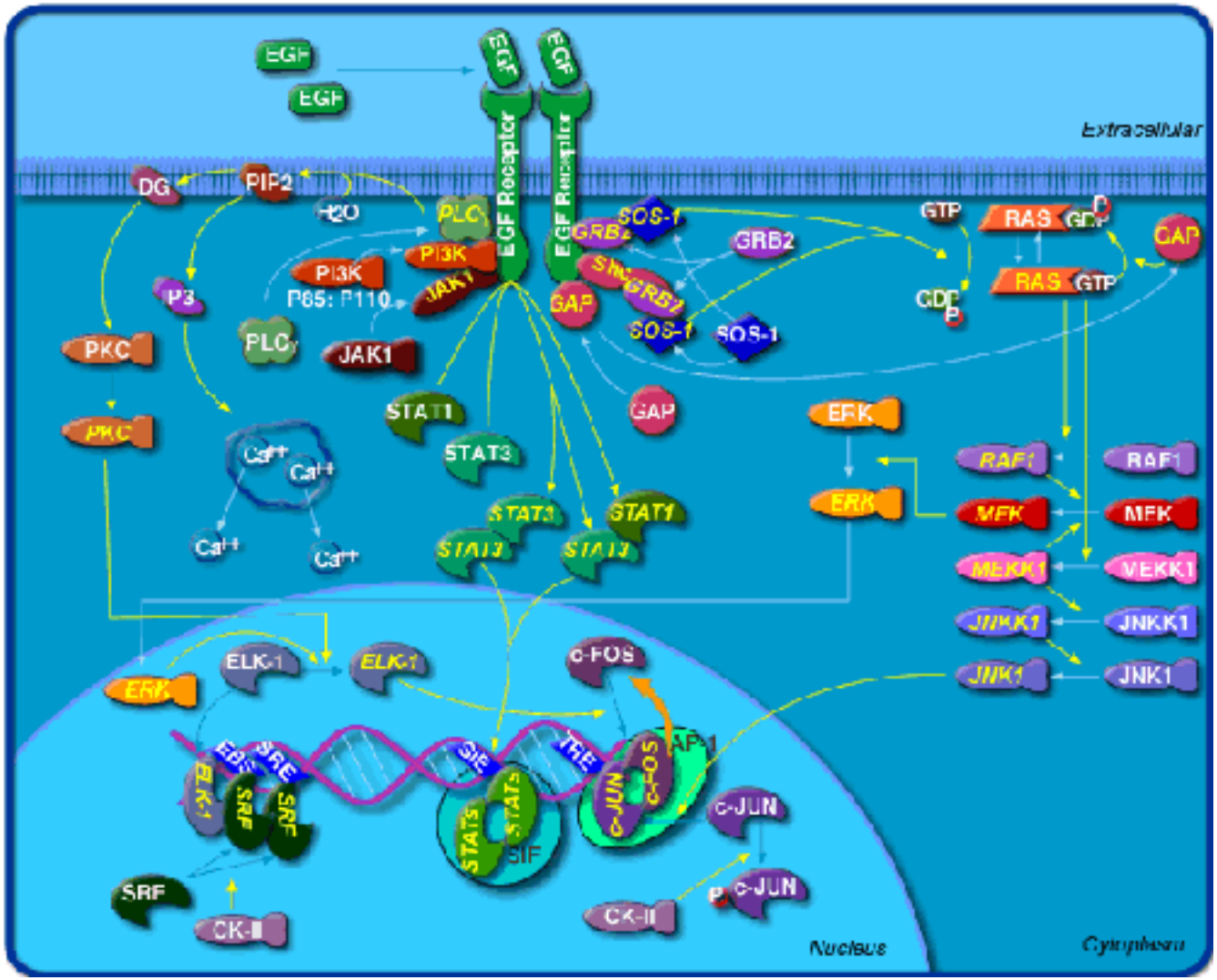
First	U	C	A	G	Last
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop (Ochre)	Stop (Umber)	A
	Leu	Ser	Stop (Amber)	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Gene finding

- The amino-acid sequence of a polypeptide is determined by the RNA sequence expressed by a gene
- The RNA sequence is determined by the gene's DNA sequence

By the central dogma, we should be able to determine the function of a protein from the DNA sequence. In practice, we can't even find the gene boundarys with 100% accuracy.

Functional genomics



The transcription of DNA to RNA is controlled by *transcription factors*, proteins that interact with the genomic DNA. These transcription factors are modulated by signal-transduction cascades. The interaction of proteins to control the expression of genes is called functional genomics, and can be studied using microarrays and gene disruption.

Evolutionary biology

Gene sequences and entire genomes show similarity from one species to another. The similarities and differences can show what parts of the genome are important.

Practical limitations

Although the central dogma states that all biological information is encoded in the DNA, we don't completely know how to decode it.

- we don't know what controls transcription (reverse-engineering genetic networks)
- we cannot completely predict RNA sequence from genomic DNA (gene finding, alternative splicing, RNA editing)
- we cannot predict protein sequence from RNA sequence (post-translational modification)
- we cannot determine protein structure from sequence (protein folding)
- we cannot determine function from structure (QSAR)

- other chemicals interact with proteins
- the environment plays a role

Sequence analysis

- we can find a sequence (or regular expression) in a set of sequences (sequence searches, gene finding)
- we can compare two sequences and determine how alike they are, and in which parts (sequence alignment)
- we can compare many sequences, and align them into matching blocks (multiple alignment)
- we can examine a multiple sequence alignment and infer evolutionary distances between the sequences (molecular phylogeny)
- we can compare many structurally similar proteins, and determine which parts of sequence determine the structure (structural alignment)

- we can determine the secondary structure of DNA, RNA and polypeptides